COMP90042

Web search and text analysis

Workshop Week 3

Your tutor

- Winn Chow (Senior Tutor)
- winn.chow1@unimelb.edu.au
- Office: Doug McDonell - 9.23
- Here, you can find my workshop slides:
- https://github.com/winnchow/COMP90042-Workshops

Postings list

Inverted Index - Recap

term t	f_t	Postings list for t
and	6	$\langle 1,6,7,8,9,12\rangle$, $\langle 1,2,1,3,1,2\rangle$
big	3	$\langle 2,5,42\rangle$, $\langle 1,1,1\rangle$
old	1	$\langle 32\rangle$, $\langle 4\rangle$
in	7	$\langle 2,3,5,6,8,14,25\rangle$, $\langle 1,1,4,1,5,3,1\rangle$
the	52	$\langle 1,2,3,4,5,7,8,9,\ldots\rangle$, $\langle 10,21,10,42,12,14,12,4,\ldots\rangle$
night	4	$\langle 1,12,13,14\rangle$, $\langle 2,2,1,3\rangle$
house	5	$\langle 6,21,32,33,43\rangle$, $\langle 2,3,4,2,1\rangle$
sleep	3	$\langle 1,51,53\rangle$, $\langle 1,2,3\rangle$
where	4	$\langle 1,3,4,6\rangle$, $\langle 1,1,2,1\rangle$

Document frequency

Document IDs

Term frequency

Compression

- How should we compress the document IDs?

term t	f_t	Postings list for t
and	6	$\langle 1, 6, 7, 8, 9, 12 \rangle$, $\langle 1, 2, 1, 3, 1, 2 \rangle$

- Document IDs: <1, 6, 7, 8, 9, 12>
- **Gaps**: <1, 5, 1, 1, 3>, so *mostly small numbers*
- Variable Byte (Vbyte) Compression

Variable Byte Compression

- For example, 1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,3
- We may encode 1,2,3 using 2 bits each.
- How about we use 0 => 1, 10 => 2, 11 => 3?

Variable Byte Compression

Examples

Number	Encoding	
824	00111000	10000110
5	10000101	

824 = 110 0111000

Storage Cost

Number Range	Number of Bytes
$0 - 127$	1
$128 - 16383$	2
$16384 - 2097151$	3

Q1 (c)

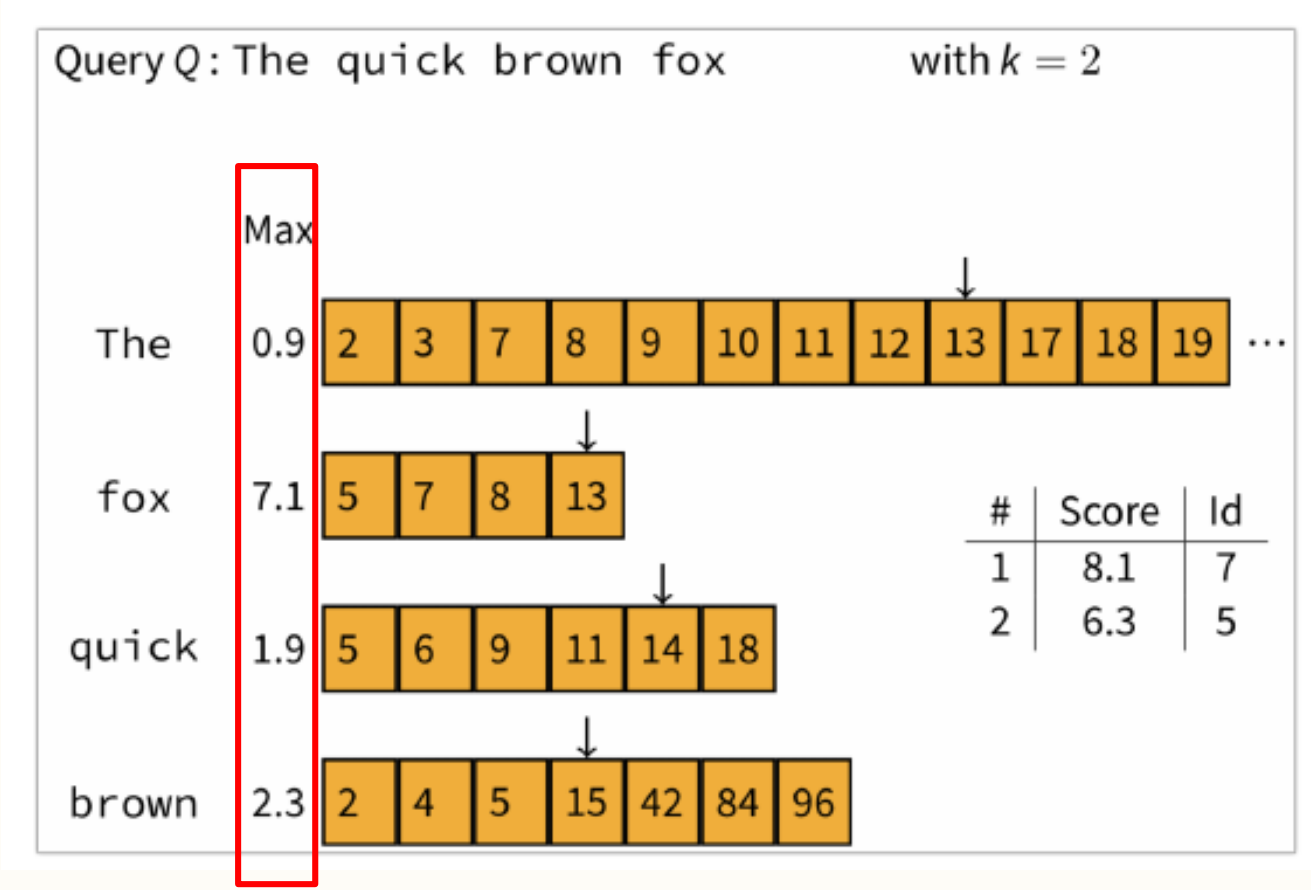
- Determine the values of integers X and Y that were encoded as the byte sequence [52,34,147,42,197] using the Variable Byte algorithm described in the lecture slides 9/10.

Q1 (c)

- 52 = 00110100
- 34 = 00100010
- 147 = 10010011
- => 0010011 0100010 0110100 = 315700
- 42 = 00101010
- 167 = 11000101
- => 1000101 0101010 = 8874

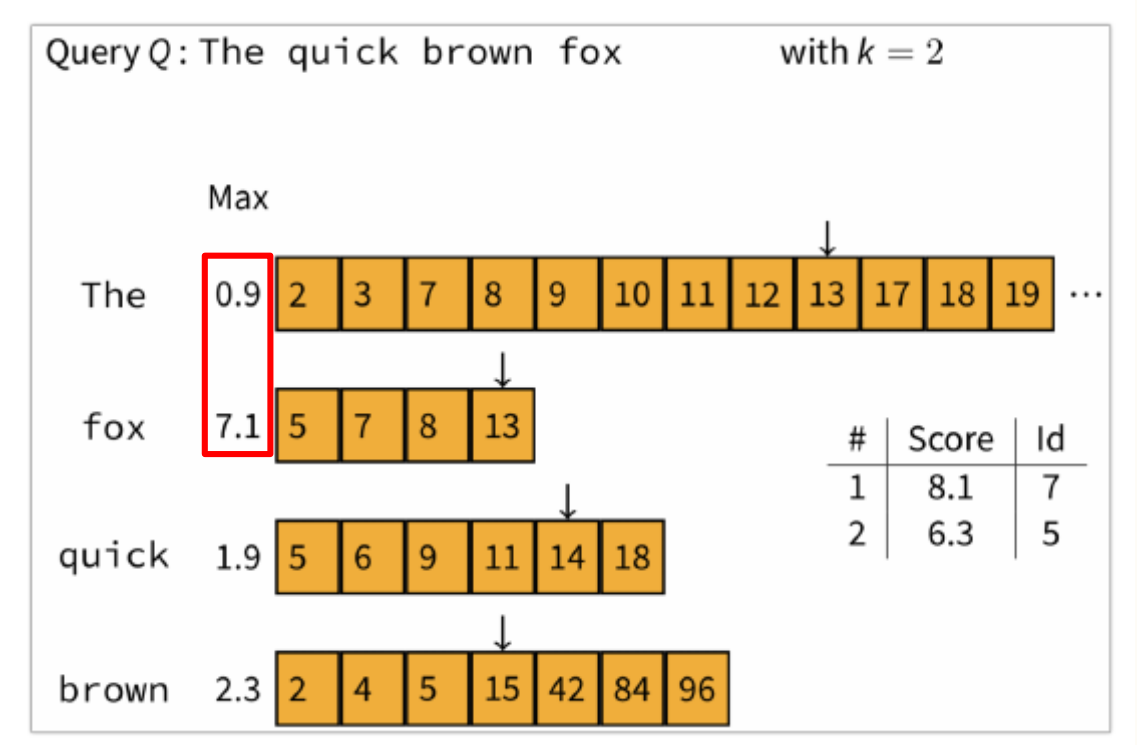
WAND – top-K query processing algorithm

$$S_{\text{TF-IDF}}(d, Q) = \sum_{t \in Q} tf_{d,t} \times \log \frac{N}{df_t}$$



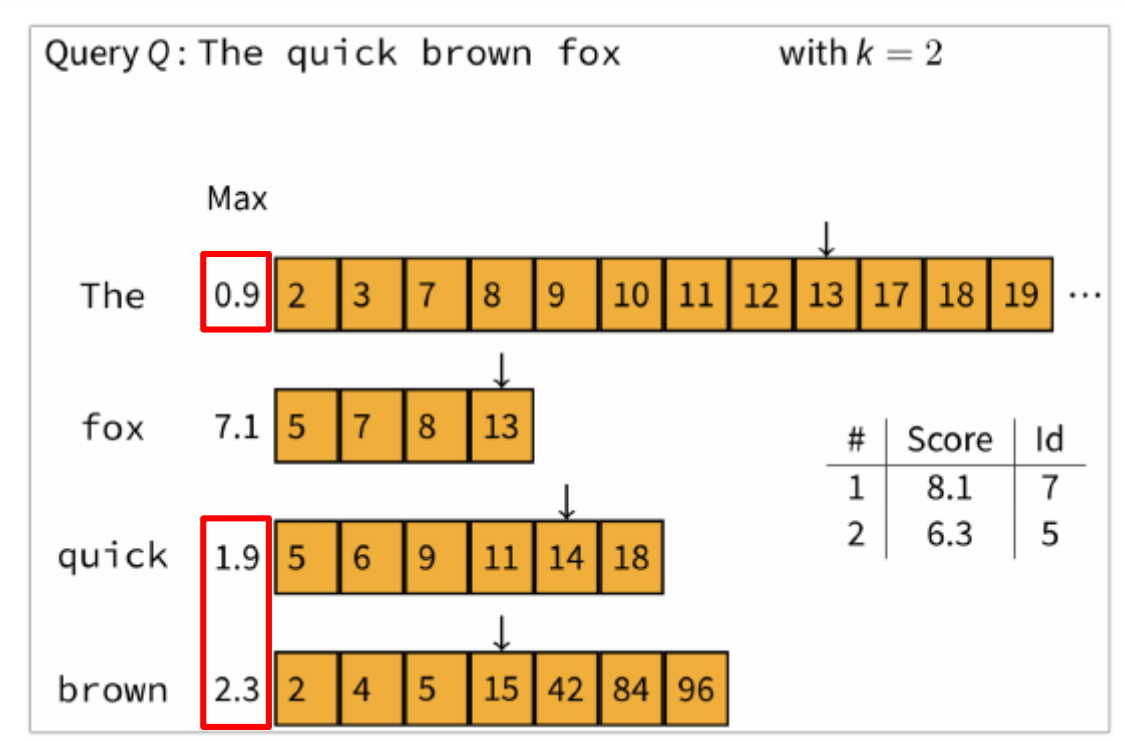
Q2

- Doc 13 is evaluated
- Max score for Doc 13 is 0.9 + 7.1 = 8.0
- So Doc 13 might enter the top-2 list



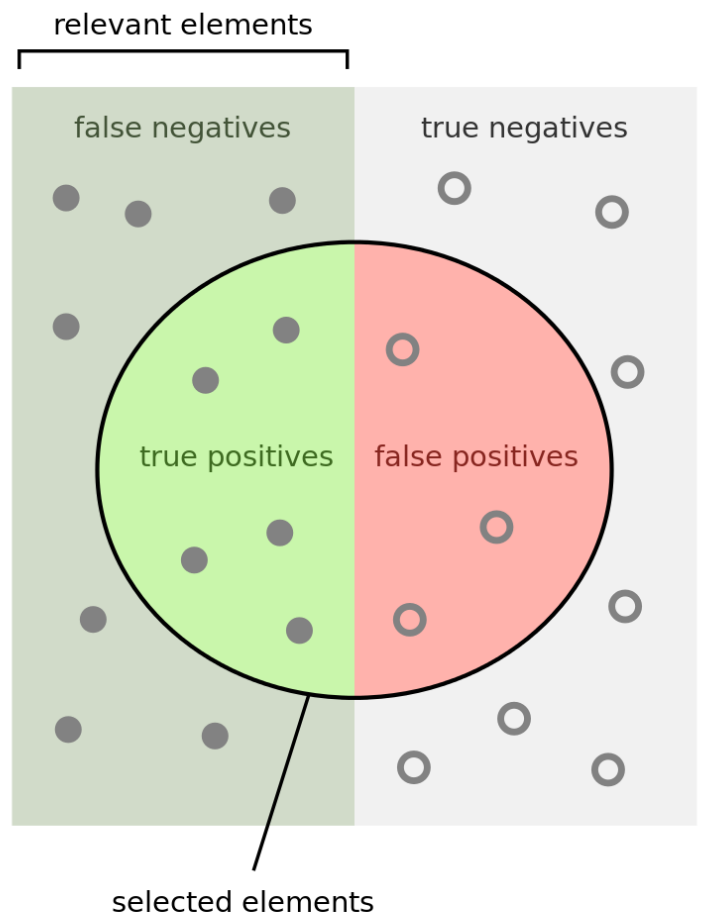
Q2

- No more documents will have “fox”
- Max score possible for a document with “The”, “quick” and “brown” will be 0.9 + 1.9 + 2.3 = 5.1
- Lower than the scores of the top-2 documents
- So, we stop.



Recall and Precision

- https://en.wikipedia.org/wiki/Precision_and_recall



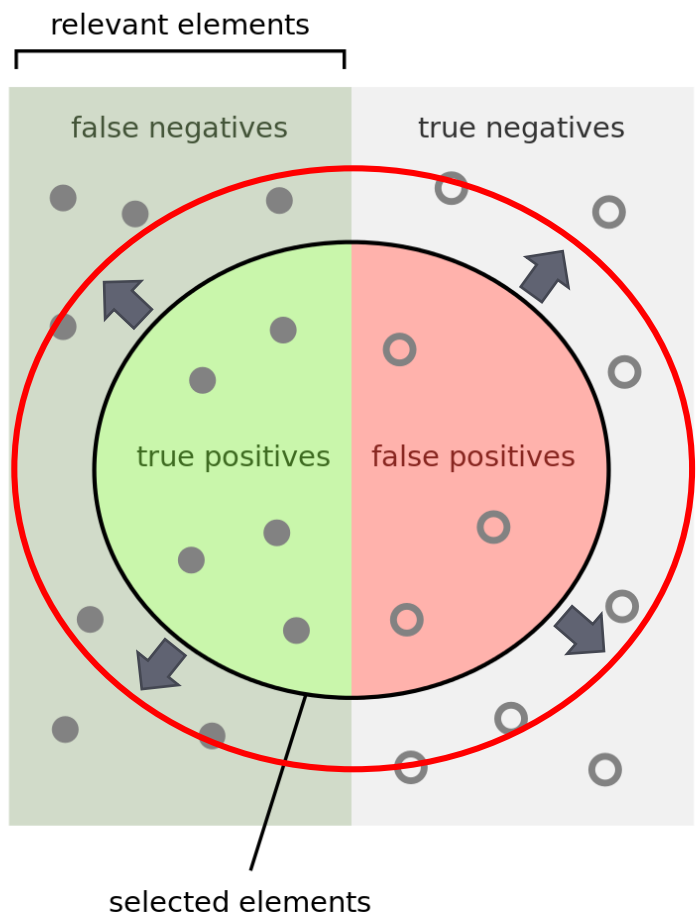
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Query Expansion



How many selected items are relevant?

Precision =

How many relevant items are selected?

Recall =

Q4

- (a) User relevance feedback
  - e.g. ask users to click
- (b) Pseudo relevance feedback
  - e.g. blind feedback, search the top-K documents and perform topic modeling
- (c) Indirect relevance feedback
  - e.g. analyze query click logs to re-rank documents

Very Useful Online Resources

- Andrei Broder - WAND Revisited
  - https://youtu.be/gwsWUPVtt6Q?t=433